Apache OpenNLP

Brochier Aymeric veille technologique
20/10/17

Contexte GITHUB

- 16 contributeurs
- 1768 commits (Last 10 days ago)
- 429 stars

Contexte : CLOC

```
aym@aym-VirtualBox:~/demo_openNPL$ cloc opennlp/
    1122 text files.
    1118 unique files.
     129 files ignored.

http://cloc.sourceforge.net v 1.60  T=4.29 s (240.3 files/s, 41900.8 lines/s)
-------------------------------------------------------------------------------
Language                     files          blank        comment           code
-------------------------------------------------------------------------------
Java                           918          16761          30898          66919
XML                             81           1220           1322          60509
Maven                            7            113            118            890
HTML                            11             56            176            150
DOS Batch                        5             25            101             65
CSS                              1             11             21             40
Bourne Shell                     5             19             92             34
YAML                             1              6              2             18
XSLT                             1              7             18              5
-------------------------------------------------------------------------------
SUM:                          1030          18218          32748         128630
-------------------------------------------------------------------------------
aym@aym-VirtualBox:~/demo_openNPL$
```

Goal

- Traitement automatique sur des textes en langage naturel

Comparaison d'outils

System	Programming Language	Target Text	Tokenization	PoS Tagging	Chunking	NER
NLTK	Python	Generic	✓	✓	✓	✓
OpenNLP	Java	Generic	✓	✓	✓	✓
CoreNLP	Java	Generic	✓	✓	✗	✓
Pattern	Python	Generic	✓	✓	✓	✗
TweetNLP	Java	Social Media	✓	✓	✗	✗
TwitterNLP	Python	Social Media	✓	✓	✓	✓
TwitIE	Java	Social Media	✓	✓	✗	✓

NER (Name Entity Recognition)

- Name Entity Recognition consiste à classifier dans un texte les mots correspondant à une certaine catégorie (les noms des personnes, les lieux, les organisations ,les dates).

Précisons et rappels

- Précision : La précision donne le pourcentage de réponses correctes
- Rappel : Le rappel donne le pourcentage des réponses correctes qui sont données

Questions pour vous

- Comment avoir un rappel 100 % ?
- Comment avoir une précision de 100 % ?

F-Mesure

Formulation [edit]

The traditional F-measure or balanced F-score (**F_1 score**) is the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Comparaison des performances

Task	Tokenization					
Data set	CoNLL			Alan Ritter - Twitter		
Metric / Tool	P $\pm \sigma$	R $\pm \sigma$	F1 $\pm \sigma$	P $\pm \sigma$	R $\pm \sigma$	F1 $\pm \sigma$
NLTK	0.95 $\pm$ 0.11	0.96$\pm$ 0.10	0.95 $\pm$ 0.11	0.83 $\pm$ 0.14	0.91 $\pm$ 0.09	0.87 $\pm$ 0.12
OpenNLP	0.99 $\pm$ 0.02	0.99 $\pm$ 0.01	0.99 $\pm$ 0.02	0.92 $\pm$ 0.11	0.96 $\pm$ 0.06	0.94 $\pm$ 0.08
CoreNLP	0.73 $\pm$ 0.31	0.73 $\pm$ 0.31	0.73 $\pm$ 0.31	0.93 $\pm$ 0.13	0.95 $\pm$ 0.11	0.94 $\pm$ 0.12
Pattern	0.42 $\pm$ 0.30	0.41 $\pm$ 0.29	0.42 $\pm$ 0.29	0.76 $\pm$ 0.21	0.78 $\pm$ 0.20	0.77 $\pm$ 0.20
TweetNLP	0.97$\pm$ 0.05	0.98 $\pm$ 0.02	0.98 $\pm$ 0.04	0.96 $\pm$ 0.07	0.98 $\pm$ 0.05	0.97 $\pm$ 0.06
TwitterNLP	0.95 $\pm$ 0.10	0.97 $\pm$ 0.09	0.96 $\pm$ 0.10	0.96 $\pm$ 0.07	0.97 $\pm$ 0.05	0.96 $\pm$ 0.06
TwitIE	0.85 $\pm$ 0.15	0.93 $\pm$ 0.11	0.89 $\pm$ 0.14	0.83 $\pm$ 0.16	0.89 $\pm$ 0.11	0.86 $\pm$ 0.13

Comparaison des performances

Task	NER					
Data set	CoNLL			Alan Ritter - Twitter		
Tool \ Metric	P $\pm \sigma$	R $\pm \sigma$	F1 $\pm \sigma$	P $\pm \sigma$	R $\pm \sigma$	F1 $\pm \sigma$
NLTK	0.88 ± 0.12	0.89 ± 0.11	0.89 ± 0.11	0.77 ± 0.16	0.84 ± 0.13	0.80 ± 0.15
OpenNLP	0.88 ± 0.09	0.88 ± 0.08	0.88 ± 0.08	0.85 ± 0.14	0.90 ± 0.11	0.87 ± 0.12
CoreNLP	0.70 ± 0.30	0.70 ± 0.30	0.70 ± 0.30	0.87 ± 0.15	0.89 ± 0.14	0.88 ± 0.15
Pattern	n/a	n/a	n/a	n/a	n/a	n/a
TweetNLP	n/a	n/a	n/a	n/a	n/a	n/a
TwitterNLP	0.88 ± 0.11	0.89 ± 0.10	0.88 ± 0.11	0.96 ± 0.07	0.97 ± 0.05	0.97 ± 0.06
TwitIE	0.74 ± 0.16	0.80 ± 0.14	0.77 ± 0.15	0.77 ± 0.17	0.83 ± 0.14	0.80 ± 0.15

Comparaison des performances

Task	PoS Tagging					
Data set	CoNLL			Alan Ritter - Twitter		
Metric / Tool	P $\pm \sigma$	R $\pm \sigma$	F1 $\pm \sigma$	P $\pm \sigma$	R $\pm \sigma$	F1 $\pm \sigma$
NLTK	0.65 $\pm$ 0.19	0.71 $\pm$ 0.18	0.68 $\pm$ 0.18	0.65 $\pm$ 0.19	0.71 $\pm$ 0.18	0.68 $\pm$ 0.18
OpenNLP	0.88 $\pm$ 0.10	0.88 $\pm$ 0.09	0.88 $\pm$ 0.10	0.70 $\pm$ 0.18	0.73 $\pm$ 0.17	0.71 $\pm$ 0.17
CoreNLP	0.67 $\pm$ 0.29	0.67 $\pm$ 0.29	0.67 $\pm$ 0.29	0.70 $\pm$ 0.19	0.71 $\pm$ 0.18	0.71 $\pm$ 0.18
Pattern	0.36 $\pm$ 0.24	0.35 $\pm$ 0.24	0.35 $\pm$ 0.24	0.61 $\pm$ 0.21	0.62 $\pm$ 0.21	0.61 $\pm$ 0.20
TweetNLP	0.83 $\pm$ 0.10	0.84 $\pm$ 0.09	0.84 $\pm$ 0.09	0.94 $\pm$ 0.08	0.96 $\pm$ 0.06	0.95 $\pm$ 0.07
TwitterNLP	0.83 $\pm$ 0.15	0.84 $\pm$ 0.15	0.83 $\pm$ 0.15	0.92 $\pm$ 0.11	0.93 $\pm$ 0.11	0.92 $\pm$ 0.11
TwitIE	0.78 $\pm$ 0.16	0.85 $\pm$ 0.12	0.82 $\pm$ 0.14	0.78 $\pm$ 0.17	0.84 $\pm$ 0.13	0.81 $\pm$ 0.14

Déduction

- En résumé si vous souhaité faire du traitement automatique que sur des tweet OpenNLP n’est pas le meilleur choix.

Démo en image : docker run

aym@aym-VirtualBox: ~/demo_openNPL

```
aym@aym-VirtualBox:~/demo_openNPL$ docker run -it joeferner/node-java bash
```

Download

```
Terminal  Fichier  Édition  Affichage  Rechercher  Terminal  Aide
aym@aym-VirtualBox:~/demo_openNPL$ sudo docker run -it joeferner/node-java bash
[sudo] Mot de passe de aym :
Unable to find image 'joeferner/node-java:latest' locally
latest: Pulling from joeferner/node-java
04c460fac791: Downloading [=====>                                          ]  6.897MB/65.77MB
0a0916b29f3e: Download complete
b25f4e7a7766: Download complete
a3ed95caeb02: Download complete
201f9ac5445c: Downloading [=>                                              ]  677.5kB/21.22MB
f9740e3107d4: Downloading [>                                               ]    528kB/71.73MB
2de2dfb54788: Waiting
7f746de2ded3: Waiting
```

Npm install java + test

test.js (~/demo_openNPL) - gedit

Ouvrir

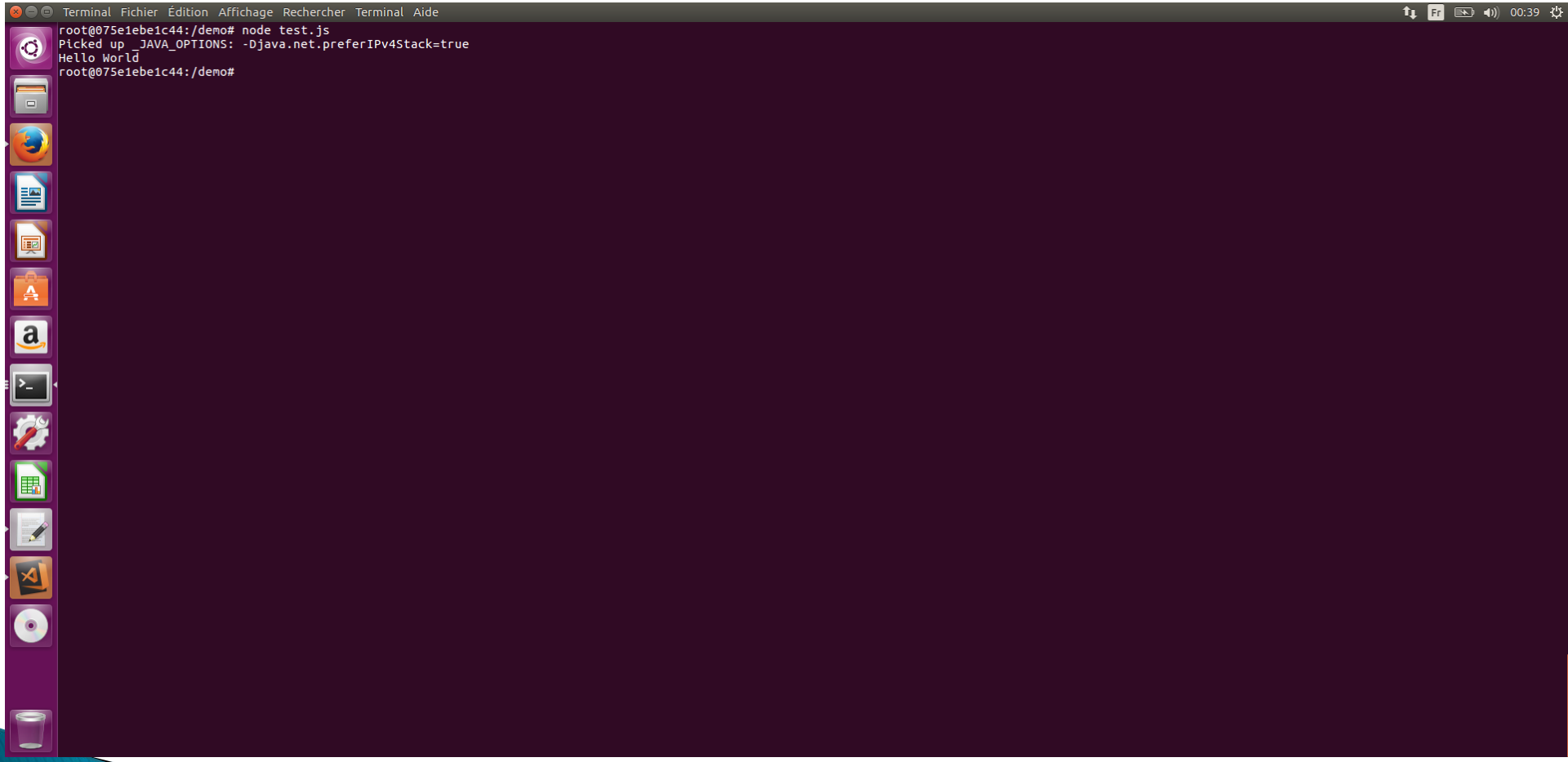
Enregistrer

remarques_utiles | demo.txt | test.js | sentenceDetector.js

```
var java = require('java');
var javaLangSystem = java.import('java.lang.System');
javaLangSystem.out.printlnSync('Hello World');
```

JavaScript | Largeur des tabulations : 8 | Lig 1, Col 1 | INS

Hello world


Terminal Fichier Édition Affichage Rechercher Terminal Aide
00:39
root@075e1ebe1c44:/demo# node test.js
Picked up _JAVA_OPTIONS: -Djava.net.preferIPv4Stack=true
Hello World
root@075e1ebe1c44:/demo#

Npm install opennlp +exemple

```
root@075e1ebe1c44:/demo# node sentenceDetector.js
Picked up _JAVA_OPTIONS: -Djava.net.preferIPv4Stack=true
[ 'Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.' ]
root@075e1ebe1c44:/demo#
```

Exemple 2 SentenceDetector

```
root@075e1ebe1c44:/demo# node sentenceDetector2.js
Picked up _JAVA_OPTIONS: -Djava.net.preferIPv4Stack=true
[ 'Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.',
  'I will try something unusual : my phone number is 06.07.08.09.10 and today M. Plum said no.' ]
root@075e1ebe1c44:/demo#
```

Exemple Tokenizer

```
root@075e1ebe1c44:/demo# node tokenizer.js
Picked up _JAVA_OPTIONS: -Djava.net.preferIPv4Stack=true
undefined [ 'Pierre',
  'Vinken',
  ',',
  '61',
  'years',
  'old',
  ',',
  'will',
  'join',
  'the',
  'board',
  'as',
  'a',
  'nonexecutive',
  'director',
  'Nov.29',
  '.' ]
root@075e1ebe1c44:/demo#
```

Exemple Tokenizer 2

```
root@075e1ebe1c44: /demo
root@075e1ebe1c44:/demo# node tokenizer2.js
Picked up _JAVA_OPTIONS: -Djava.net.preferIPv4Stack=true
undefined [ 'Pierre',
  'Vinken',
  ',',
  '61',
  'years',
  'old',
  ',',
  'will',
  'join',
  'the',
  'board',
  'as',
  'a',
  'nonexecutive',
  'director',
  'Nov.29',
  '.',
  'I',
  'found',
  'this',
  '06.07.08.09.10',
  '.',
  'What',
  'is',
  'it',
  '?' ]
root@075e1ebe1c44:/demo#
```

Questions